Anal. Chem. **2010,** *82,* 4386–4395

Metabolite Differentiation and Discovery Lab (MeDDL): A New Tool for Biomarker Discovery and Mass Spectral Visualization

Claude C. Grigsby,*,‡,|| Mateen M. Rizki,‡ Louis A. Tamburino,‡ Rhonda L. Pitsch,§ Pavel A. Shiyanov,§ and David. R. Cool||

Counter Proliferation Branch, Biosciences and Protection Division, Human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson AFB, Dayton, Ohio 45433-5707; Department of Computer Science and Engineering, Wright State University, 3640 Colonel Glenn Highway, Dayton, Ohio 45435; Applied Biotechnology Branch, Biosciences and Protection Division, Human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson AFB, Dayton, Ohio 45433-5707; and Department of Pharmacology and Toxicology, Boonshoft School of Medicine, Wright State University, 3640 Colonel Glenn Highway, Dayton, Ohio 45435

The goal of this work was to design and implement a prototype software tool for the visualization and analysis of small molecule metabolite GC−MS and LC−MS data for biomarker discovery. The key features of the Metabolite Differentiation and Discovery Lab (MeDDL) software platform include support for the manipulation of large data sets, tools to provide a multifaceted view of the individual experimental results, and a software architecture amenable to modification and addition of new algorithms and software components. The MeDDL tool, through its emphasis on visualization, provides unique opportunities by combining the following: easy use of both GC−MS and LC−MS data; use of both manufacturer specific data files as well as netCDF (network Common Data Form); preprocessing (peak registration and alignment in both time and mass); powerful visualization tools; and built in data analysis functionality.

Metabolomics is a rapidly growing field used to characterize the metabolic profile of a specific tissue or biofluid. Metabolic profiling, originally pioneered by Jeremy Nicholson, Elaine Holmes, and John Lindon at the Imperial College in London[1] utilizing nuclear magnetic resonance (NMR) based analysis, has evolved to become one of the most common applications of liquid chromatography mass spectrometry (LC−MS).[2−4] Metabolomics is an attractive approach to the study of time-related quantitative multivariate metabolic responses to pathophysiological processes by which biological and chemical agents, e.g., drugs, can cause perturbations in the concentrations and flux of endogenous metabolites involved in critical cellular pathways.[5] Thus, cells respond to toxic insult or other stressors by altering their intra- and/or extra-cellular environment in an attempt to maintain a homeostatic intracellular environment.

This metabolic alteration is expressed as a "fingerprint" of biochemical perturbations characteristic of the type and target of a toxic insult or disease process.[6] These metabolic alterations are often seen in body fluids as changes in metabolic profiles in response to toxicity or disease, as the body attempts to maintain homeostasis by eliminating substances from the body. Therefore, because many biofluids can be easily obtained either noninvasively (urine) or minimally invasively (blood), they are typically used in metabolomic studies.[7] Additionally, if a significant number of trace molecules can be identified and monitored, the overall pattern produced may be more consistent and predictive than any single biomarker,[8] which would prove of great value in the development of deployable devices for testing toxic or infectious exposures.

Current LC−MS systems typically consist of a system of specialized instrumentation with customized support software. This software is generally proprietary, being supplied by the instrument manufacturer and designed to facilitate user interaction with the analytical hardware. Most LC−MS manufacturers also market add-on commercial software packages for the analysis of the results of LC−MS experiments, which are generally designed to provide a very specific type of data analysis (i.e., proteomic or metabolomic) and cannot be readily modified or added to by the end-user. For larger metabolomic biomarker discovery studies, such as the LC−MS effort initiated by our laboratory for profiling low level exposure to environmental toxicants, none of the software solutions available at the time offered the ability to compare multiple time point and dosage groups, or handle data sets in significant sample numbers (>100 samples in duplicate). This bottleneck in data handling initiated the development of the

* To whom correspondence should be addressed. Fax: (937)558-8474. E-mail: claude.grigsby@wpafb.af.mil.

‡ Counter Proliferation Branch, Biosciences and Protection Division, Human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson AFB.

|| Department of Pharmacology and Toxicology, Boonshoft School of Medicine, Wright State University.

§ Applied Biotechnology Branch, Biosciences and Protection Division, Human Effectiveness Directorate, Air Force Research Laboratory, Wright-Patterson AFB.

(1) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181–1189.

(2) Pham-Tuan, H.; Kaskavelis, L.; Daykin, C.; Janssen, H. *J. Chromatogr. B* **2003**, *789*, 283–301.

(3) Lindon, J.; Holmes, E.; Nicholson, J. *Pharm. Res.* **2006**, *23*, 1075–1088.

(4) Robertson, D. *Toxicol. Sci.* **2005**, *85*, 809–822.

(5) Nicholson, J.; Connelly, J.; Lindon, J.; Holmes, E. *Nat. Rev. Drug Discov.* **2002**, *1*, 153–161.

(6) Reo, N. *Drug Chem. Toxicol.* **2002**, *25*, 375–382.

(7) Dunn, W.; Ellis, D. *Trends Anal. Chem.* **2005**, *24*, 285–294.

(8) Wang, Y.; Holmes, E.; Nicholson, J.; Cloarec, O.; Chollet, J.; Tanner, M.; Singer, B.; Utzinger, J. *Proc. Natl. Acad. Sci.* **2004**, *101*, 12676–12681.

10.1021/ac100034u
Published on Web 05/05/2010

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 JUN 2010		

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Metabolite Differentiation and Discovery Lab	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **711 HPW/RHPC,2729 R Street,Bldg. 837 Area B,Wright-Patterson AFB ,OH,45433-5707**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goal of this work was to design and implement a prototype software tool for the visualization and analysis of small molecule metabolite GC-MS and LC-MS data for biomarker discovery. The key features of the Metabolite Differentiation and Discovery Lab (MeDDL) software platform include support for the manipulation of large data sets, tools to provide a multifaceted view of the individual experimental results, and a software architecture amenable to modification and addition of new algorithms and software components. The MeDDL tool, through its emphasis on visualization, provides unique opportunities by combining the following: easy use of both GC-MS and LC-MS data; use of both manufacturer specific data.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**10**	

MeDDL tool described below, allowing us to differentiate metabolite profiles in a large time-dose study and facilitating the ability to visualize exposure data for a global view of an entire chemical-exposure set while maintaining the ability to focus on individual metabolites and spectra for subsequent identification.

To illustrate the novel visualization and rapid multisample analysis capability of MeDDL for discovery of metabolomic biomarkers, data from a portion of this study examining environmental toxicant exposure have been selected. Environmental exposures to toxins as well as therapeutic interventions often cause nephrotoxicity.[9] An expanded list of metabolites indicating kidney damage would be immensely helpful in the monitoring of renal conditions after exposure to external toxicants, not only in pharmaceutical drug safety evaluations and clinical studies, but also in the occupational and military operational setting.

As reported previously by our group,[10] D-serine is ubiquitous in human plasma and comprises up to 3% of total plasma serine level in humans, with plasma D-serine elevations observed in chronic renal failure, suggesting elimination by the kidney is responsible for control of D-serine concentrations. D-serine is reabsorbed in the pars recta region of the rat proximal tubule and subsequently metabolized by D-aminoacid oxidase (D-AAO), to produce α-keto acid, ammonia, and hydrogen peroxide.[11,12] Other research has indicated that metabolism of D-serine by D-AAO is causative for initiation of toxicity in the kidney, with elevated levels generating selective necrosis to the pars recta region of the renal proximal tubules in the rat.[13] The choice to use the D-Serine model was made in order to reveal both early and sensitive biomarkers for epithelial cell injury in the kidney.

EXPERIMENTAL SECTION

Urine Samples and Materials. Animal use in this study was conducted in accordance with the principles stated in the Guide for the Care and Use of Laboratory Animals, National Research Council, 1996, and the Animal Welfare Act of 1966, as amended. Male Fischer 344 rats weighing 222−258 g were obtained from Charles River Laboratories. Groups of five animals received a single intraperitoneal (IP) dose of D-serine at a dose of 5, 20, or 500 mg/kg (or vehicle only −0.9% saline solution). Food (Purina Certified Rat Chow #5002) and water was available for all animals ad libitum. The housing environment was maintained on a 12 h light−darkness cycle at 25 °C, and all animals were examined by Vivarium personnel twice daily. Urine samples were collected cold using plastic 50 mL conical tubes containing 1.0 mL of 1% sodium azide maintained at 6−10 °C using I-Cups (Bioanalytical Systems, Inc.; stored at −80 °C prior to use) 24 h prior to dosing and daily thereafter, generating five 24-h intervals (0, 24, 48, 72, and 96 h postdosing). The urine was then frozen at −20 °C and thawed on ice prior to analysis. For the D-serine exposure set described, 104 individual samples were processed by aliquoting 1.0 mL of urine into a 2 mL centrifuge tube and centrifuged at 13 000 rpm for 5 min at 5 °C to remove debris. The supernatant was removed using a 1 mL tuberculin syringe and filtered through a 0.2 μm PTFE syringe filter disk prior to aliquot transfer to two Waters Corp. Total Recovery Vials and subsequent duplicate testing.

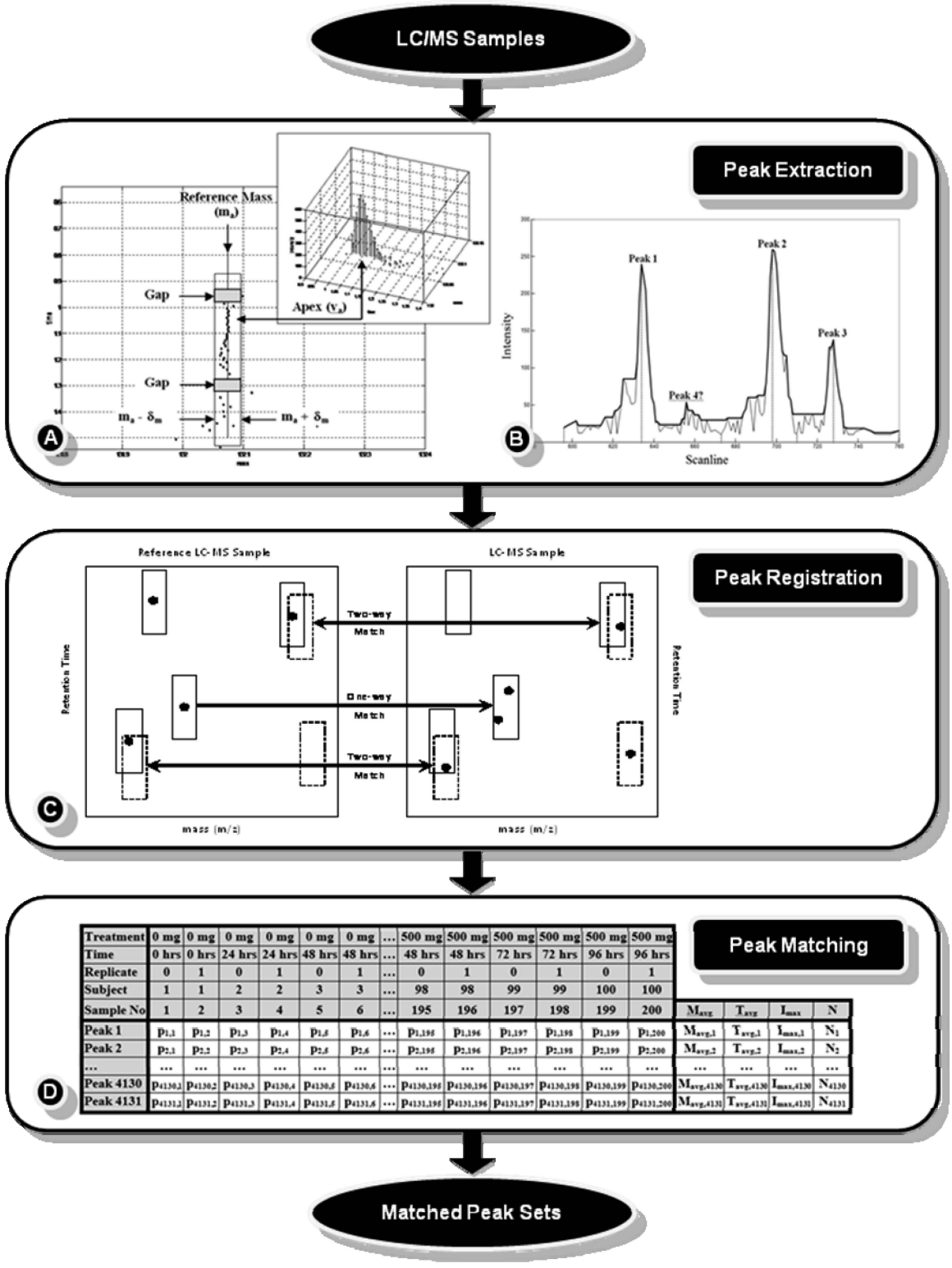
Instrumentation and Methods. The LC−MS system utilized for sample analysis was a Waters Q-ToF Micro in line with a Waters Acquity UPLC. The source temperature was set to 130 °C, a desolvation gas temperature of 320 °C and a desolvation gas flow of 600 L/h were employed. The capillary voltage was set at 3.2 kV for both positive and negative ion mode analysis. A scan time of 0.4 s with an interscan delay of 0.1 s was used throughout, and data were collected in centroid mode. A 1-μL aliquot of filtered urine was injected onto a 2.1 × 100 mm, 1.7 μm Acquity UPLC BEH C18 column (Waters Corporation) held at 40 °C. Retained small molecules were eluted via a linear gradient of 98% A for 2 min, 2−50% B from 2−11 min, 50−98% B over 12−12.49 min, returning to 98% A at 12.5 min and remaining there until completion of the run at 15 min at an eluent flow rate of 0.25 mL/min; where A = 0.1% formic acid and B = 0.1% formic acid in acetonitrile. The mass spectrometric data were collected in full scan mode from m/z 80 to 1000 from 0.8 to 15 min. Urine samples were run in duplicate and analyzed using MeDDL using spectra from 0.8−12 min. For MS/MS data, random urine samples were run using data dependent acquisition with multiple voltages applied. Standards were purchased from Sigma-Aldrich (St. Louis, MO) and run at 1 mg/mL (1 μg injection) under the same LC−MS conditions as those of the samples to validate retention times and MS/MS spectra. Sample analysis for determination of differential metabolites was performed using the MeDDL tool which is described below.

ALGORITHMS AND IMPLEMENTATIONS

MeDDL Overview. The overall goal of the MeDDL system is to facilitate the analysis of LC−MS experimental results. With this goal in mind, the system is structured to provide a global view of experimental results so a user can quickly identify samples exhibiting interesting or unusual patterns of behavior while still having the option to probe these samples at ever finer levels of detail. MeDDL accomplishes this by allowing the user to search for relationships between subsets of subjects at selected times or treatment levels. The user may ask for subsets which exhibit specific levels of change in the behavior of the response. The user may restrict the fold-change to positive, negative or combined levels of changes. For example, the user can seek all peaks that exhibited a 5-fold positive change between the control subjects and treated subjects at 24 or 48 h. In addition, MeDDL also allows the user to perform detailed statistical analysis including ANOVA (1-way, 2-way, and N-way) among the selected subject groups. The user can optionally perform multiple pairwise comparison tests among the means of groups to determine whether or not all differences among group means satisfy a user defined level of significance. A Bonferroni correction is applied to compensate for the tendency to incorrectly find a single pairwise significant difference among multiple comparisons.

The MeDDL system is composed of two major subsystems: peak analysis and visualization. Peak analysis encompasses several subsidiary tasks including peak extraction, peak registration, and extraction of registered peaks sets. The visualization system takes the information provided by the peak analysis subsystem and combines it with information describing the overall experiment

(9) Van Vleet, T.; Schnellmann, R. 2003; Elsevier; 500−508.
(10) Soto, A.; DelRaso, N. J.; Schlager, J. J.; Chan, V. T. *Toxicology* **2008**, *243*, 177–192.
(11) Carone, F.; Ganote, C. *Arch. Pathol.* **1975**, *99*, 658.
(12) Ganote, C.; Peterson, D.; Carone, F. *Am. J. Pathol.* **1974**, *77*, 269.
(13) Pilone, M. S. *Cell. Mol. Life Sci.* **2000**, *57*, 1732–1747.



Treatment	0 mg	0 mg	0 mg	0 mg	0 mg	0 mg	…	500 mg	500 mg	500 mg	500 mg	500 mg	500 mg
Time	0 hrs	0 hrs	24 hrs	24 hrs	48 hrs	48 hrs	…	48 hrs	48 hrs	72 hrs	72 hrs	96 hrs	96 hrs
Replicate	0	1	0	1	0	1	…	0	1	0	1	0	1
Subject	1	1	2	2	3	3	…	98	98	99	99	100	100
Sample No	1	2	3	4	5	6	…	195	196	197	198	199	200
Peak 1	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$	$p_{1,4}$	$p_{1,5}$	$p_{1,6}$	…	$p_{1,195}$	$p_{1,196}$	$p_{1,197}$	$p_{1,198}$	$p_{1,199}$	$p_{1,200}$
Peak 2	$p_{2,1}$	$p_{2,2}$	$p_{2,3}$	$p_{2,4}$	$p_{2,5}$	$p_{2,6}$	…	$p_{2,195}$	$p_{2,196}$	$p_{2,197}$	$p_{2,198}$	$p_{2,199}$	$p_{2,200}$
…	…	…	…	…	…	…	…	…	…	…	…	…	…
Peak 4130	$p_{4130,1}$	$p_{4130,2}$	$p_{4130,3}$	$p_{4130,4}$	$p_{4130,5}$	$p_{4130,6}$	…	$p_{4130,195}$	$p_{4130,196}$	$p_{4130,197}$	$p_{4130,198}$	$p_{4130,199}$	$p_{4130,200}$
Peak 4131	$p_{4131,1}$	$p_{4131,2}$	$p_{4131,3}$	$p_{4131,4}$	$p_{4131,5}$	$p_{4131,6}$	…	$p_{4131,195}$	$p_{4131,196}$	$p_{4131,197}$	$p_{4131,198}$	$p_{4131,199}$	$p_{4131,200}$

M_{avg}	T_{avg}	I_{max}	N
$M_{avg,1}$	$T_{avg,1}$	$I_{max,1}$	N_1
$M_{avg,2}$	$T_{avg,2}$	$I_{max,2}$	N_2
…	…	…	…
$M_{avg,4130}$	$T_{avg,4130}$	$I_{max,4130}$	N_{4130}
$M_{avg,4131}$	$T_{avg,4131}$	$I_{max,4131}$	N_{4131}

Figure 1. Schematic overview of the peak analysis subsystem. (A) Sample peak cluster extraction. (B) Sample peak cluster partition. (C) Preliminary peak matching for registration. (D) Matched Peaks. $\boldsymbol{p_{i,j}} = <\boldsymbol{M_{i,j}}, \boldsymbol{T_{i,j}}, \boldsymbol{I_{i,j}}>$—apex point of extract peak with mass ($\boldsymbol{M_{i,j}}$), register time ($\boldsymbol{T_{i,j}}$) and intensity ($\boldsymbol{I_{i,j}}$). The behavior of each row is summarized by the average mass ($\boldsymbol{M_{avg,i}}$), average registered time ($\boldsymbol{T_{avg,i}}$), average intensity ($\boldsymbol{I_{max,i}}$) and a count of the number of peaks detected ($\boldsymbol{N_i}$).

to allow the user to explore the results from the perspective of the experimental parameters.

Peak Analysis Subsystem. In this section, we present a brief overview of the algorithms that comprise the peak analysis subsystem. As shown in Figure 1, these algorithms include peak extraction, peak registration and peak matching. Each algorithm is described below.

Peak Extraction Algorithm. The peak extraction algorithm is composed of two phases. The first phase is designed to form temporal clusters required for chromatographic time alignment

and the second phase partitions clusters into individual peaks. A full experiment consists of hundreds of files from different LC−MS sample runs, each of which is identified by subject, treatment, time, and optionally replication indices. A single LC−MS sample is composed of a set of n measurement points $P = \{p_i \mid i = 1, 2, 3, \ldots n\}$ of the form: $p_i = (M_i, L_i, I_i)$ with components of mass (M_i), scan number (L_i) and intensity value (I_i). Each scan (L_i) also has a corresponding retention time (T_i). Extracted peaks are temporal sequences of similar mass coordinates across multiple scans. A simple example of the algorithm for forming peak clusters is shown in Figure 1A.

The peak extraction process is initialized by selecting a reference point (M_a, L_a, I_a) with a large intensity which will become the apex of the resultant peak bounded by a narrow mass band. The width of the mass band is set by a mass uncertainty parameter (Δ_m) specified by the user. Within this mass-band, points are assembled into cluster sequences in both temporal directions from the initial reference point, accepting only one point per scan. A resultant cluster may grow to lengths spanning many peaks.

The next step, partitioning the cluster into individual peaks, is a difficult design problem, because it must instantiate a peak definition that separates the significant peaks from the noisy and uninteresting ones. Often, partitioning a cluster visually can be difficult, so some ambiguous results are unavoidable. In other words, if it is difficult to resolve peaks visually, it is difficult to automate. In Figure 1B there are many small, jagged, noisy peaks and three or four prominent peaks. The decision to extract 3 or 4 peaks is determined by adjusting a user-accessible control parameter. In this example, a closing operator from mathematical morphology[14] has been employed to filter out unwanted peaks, including the fourth obvious candidate. The horizontal fill-in lines are determined by the size of the structuring element used by the morphological operators to probe the cluster's structure.

Peak Registration. Peak registration uses only mass and retention time coordinates (M_a,T_a) of the peak apexes to achieve the significant data reduction required to work efficiently across many LC−MS samples. Peak registration primarily involves temporal alignment of peaks, although for some instruments the alignment of mass measurements is also required. Initially, one of the images is selected as the reference image and all others are transformed to match it. As illustrated in Figure 1C, this transformation is accomplished by bracketing each peak with a maximum shift window (Δ_m, Δ_t) and identifying matching pairs of peaks. Ideally, these are one-to-one unique matches, meaning each peak has only one unique candidate match. A larger set of candidate matched peaks can be defined by relaxing the criteria so that only the peak in either the reference image or alignment images has a unique matching peak. The set of matched peaks is then used to compute a polynomial transformation that maps retention times of images relative to the reference image. The order of the polynomial is determined by the user (eq 1).

$$T_r = \text{polyval}\,(T, \text{polyCoef}) = c_2 \cdot T^2 + c_1 \cdot T + c_0 \ldots \quad (1)$$

Should the need arise to make adjustments to mass coordinates (eq 2); the set would be used to compute a bivariate alignment polynomial.

$$T_r = \text{polyval}\,(T, M, \text{polyCoef}) \quad (2a)$$

$$M_r = \text{polyval}\,(T, M, \text{polyCoef}) \quad (2b)$$

Peak Matching. The set of matching pairs of peaks is used to initialize a matrix of matched peaks (Figure 1D). Each column represents one image and each row contains a set of registered peaks. The peak coordinates (M_a, T_a) are averaged over non-empty images in each row (M_{avg}, T_{avg}), producing a synthetic reference image so that the original reference image is no longer required. A number of cycles of the matching algorithm are then used to fill in existing rows and to extend the number of rows by seeding the reference image with peaks from the pool of unused peaks. The coordinates of the seed peaks are used as initial values for (M_{avg},T_{avg}). Each iteration of the matching algorithm produces new alignment polynomials by pairing image peak coordinates with the evolving row averages. A peak matches the row average if its coordinates fall within a (Δ_m, Δ_t) box centered on the row average, where Δ_t is much smaller than the Δ_m used for prealignment matching pairs. The final matching step, which attempts to fill in any empty slots in the matching matrix, is accomplished by selecting the raw

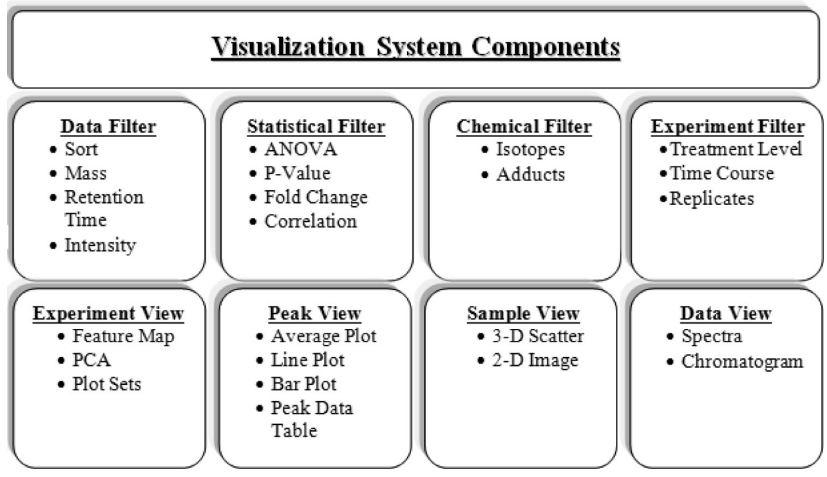


Figure 2. Visualization system overview.

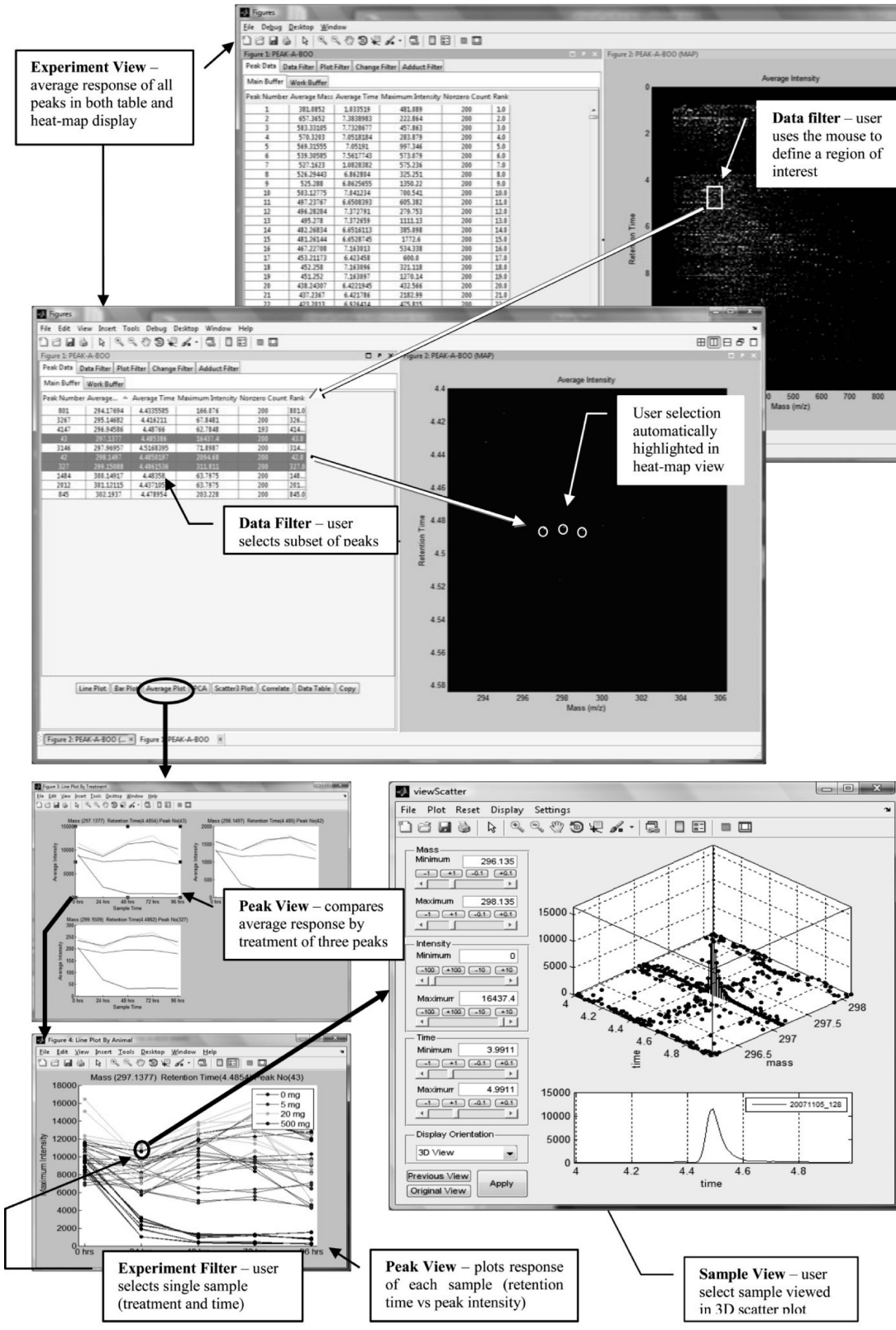


Figure 3. Example MVC (model-view-controller) interaction.

data point with the maximum intensity in the (Δ_m, Δ_t) acceptance box as a peak substitute.

Visualization System. The visualization system is based on the MVC (model-view-controller) software architecture pattern.[15] The model is composed of a series of relational data tables that include the registered match peak table (Figure 1D), the experimental descriptor table, cluster-peak data tables and raw sample data tables. The user interacts with the model via a graphical interface that supports mouse and keyboard input. The communication between the controller and the model is implemented

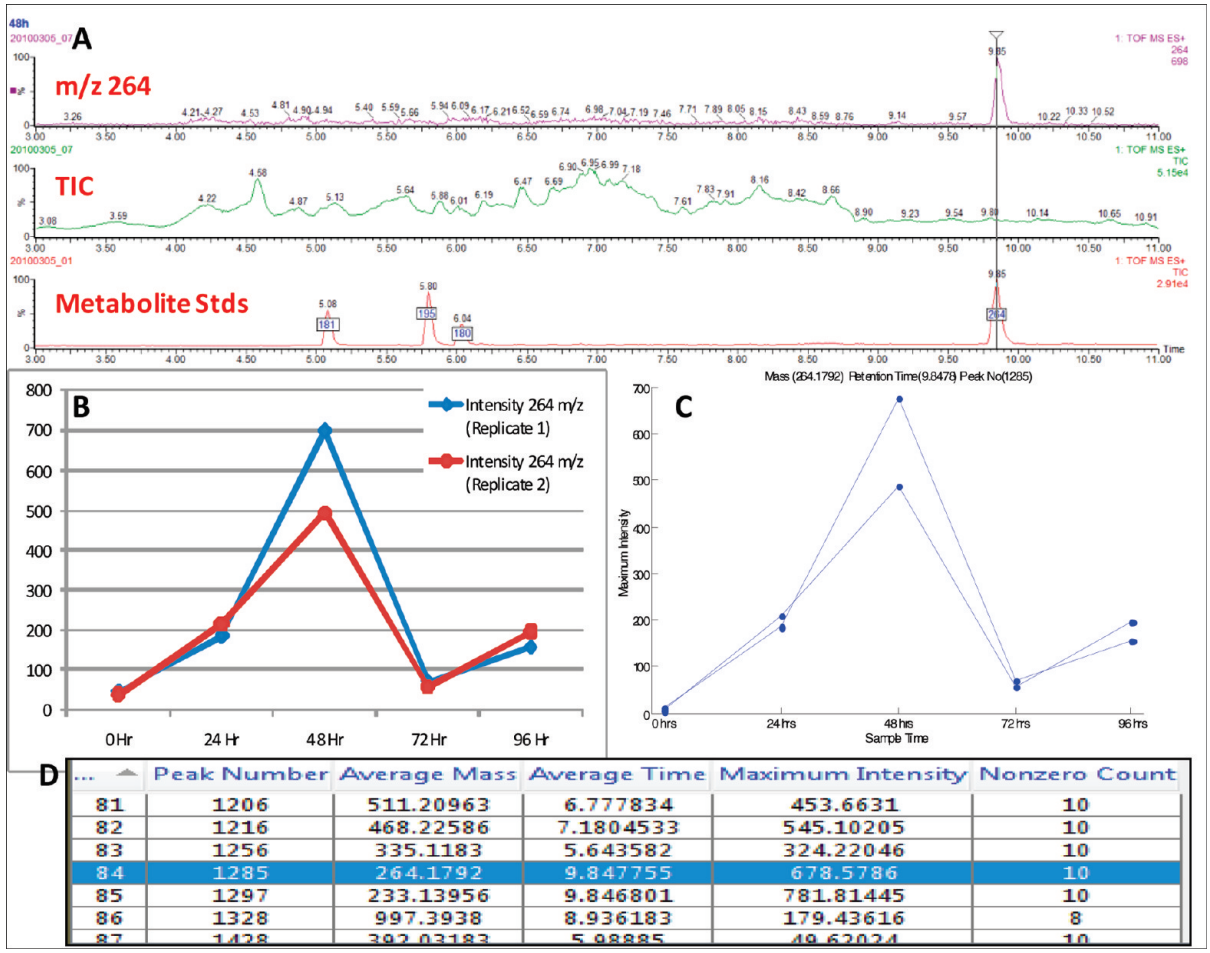


Figure 4. (A) Spectra from spiked study of control F344 rat urine showing presence of nortryptyline in urine, TIC of urine sample, and TIC of test mixture utilized in spike. (B) Nortryptyline dose response obtained via Masslynx (Waters Corporation). (C) Nortryptyline dose response obtained via MeDDL following alignment and registration. (D) Selection of nortryptyline via MeDDL as showing >5-fold change in time versus treatment.

using callback mechanisms defined in the Matlab programming language. When the user triggers a callback event, the controller notifies the model of the user's action and then possibly alters the state of the model. The view may automatically be invoked by the controller to update some subset of displays as a result of a change in the state of the model or the view may query the model to generate a display based on a user request.

The user interaction with the model is organized into a collection of filters that allow the user structured access to the various components of the data model (Figure 2). These filters are divided into logical categories: data, statistical, chemical, and experimental. The data filters allow access to subsets of data that are restricted by mass, retention time or intensity. The statistical filters allow the user to locate statistically significant patterns of behavior across the entire set of registered peaks. Chemical filters allow the user to remove certain peaks from the analysis based on chemical properties. For example, isotopic peaks or adducts can be automatically filtered to simplify analysis. Finally, experiment level filters allow the user to select items related to the biological experiment such as treatments levels or longitudinal studies for analysis.

The filtered data is visualized through a variety of displays. The displays allow a multifaceted view of the data. Figure 3 demonstrates one series of filter-display interactions possible using the visualization system. In this example, the main display opens with a view of all registered peaks stored in a summary table along with a heat-map. Each point in the heat-map represents the location of a registered set of matched peaks. The position of the point in the heat-map denotes the peak-set's average mass and average retention time. These correspond to the values of (M_{avg}, T_{avg}) shown in Figure 1D. The brightness of the point is determined by the value of the most intense peak in the registered set (I_{max} in Figure 1D). As shown in Figure 3, the user can apply a data filter to identify a smaller set of registered peaks for analysis and then alter the view to show line plots summarizing the behavior of several registered peaks. The user can select a single registered peak, plot the behavior of all samples as a function of the experimental parameters (treatment vs time), and select a specific sample for further analysis. Additionally, the user can explore the raw data in a 3D scatter plot with the ability to zoom-in and zoom-out in any specified spectral region.

The MeDDL tool is freely available and the described Matlab module will be provided upon request by the corresponding author (claude.grigsby@wpafb.af.mil) to interested parties. The platform independent Python code will be made available upon completion.

(14) Serra, J.; London: Academic Press, 1988.
(15) Buschmann, F. *Pattern-Oriented Software Architecture: A System of Patterns*; Wiley: Chichester, New York, 1996.

Table 1. List of Spiked Standards (Waters Corporation Metabolite Test Mix) Added to F344 Control Urine Utilized in Software Validation Study As a Synthetic Dose Response

	theophylline	caffeine	hippuric acid	4-nitrobenzoic acid	nortryptyline
0 h	0 pg	0 pg	0 pg	0 pg	0 pg
24 h	750 pg	750 pg	750 pg	375 pg	281 pg
48 h	3.75 ng	3.75 ng	3.75 ng	1.88 ng	1.41 ng
72 h	375 pg	375 pg	375 pg	188 pg	141 pg
96 h	750 pg	750 pg	750 pg	375 pg	281 pg

RESULTS AND DISCUSSION

To illustrate the utility and power of MeDDL in the visualization of large, multigroup experiments, we analyzed data from an LC−MS effort for profiling low level kidney biomarkers in the F344 rat model. Importation of the raw QToF MS data files and subsequent analysis of the aligned and registered peak database by MeDDL identified numerous metabolic changes in the urine of the animals after D-Serine treatment compared to control animals. Registration and processing of the D-serine exposure data ($n = 208$) utilizing peak inclusion criteria requiring each m/z at a given retention time be present in a minimum of 5% of all samples was accomplished in 122 min. A smaller sample set of $n = 20$ was similarly registered and processed in 12 min to establish scalability, with all analysis performed utilizing a dual core 2.53 GHz CPU with 6 GB of RAM. This alignment and registration encompassed all detectable peaks, with absolute intensity values as low as 30 being registered (background level previously established by our group for the QToF Micro).

To verify the accuracy of the registration and analysis algorithms, a spiked study of control F344 rat urine was performed (Figure 4A) using a purchased metabolite test mixture of five known compounds (Waters Corporation Metabolomic Test Mix). An artificial dose response was generated as shown below (Table 1) and examination of nortryptyline (m/z 264.1752), the highest intensity standard in this set, via Masslynx (Waters Corporation) generated the response illustrated in Figure 4B (below). Following processing by MeDDL, nortryptyline was registered by the software generating an identical response curve to that manually determined in the vendor supplied instrument control software (Figure 4C). Analysis of the spiked data utilizing the previously described fold change filter for all masses showing a 5-fold change across time for a given dose showed inclusion of nortryptyline (Figure 4D). Accuracy in both the ability of the software to perform correlations as well as in peak registration can be demonstrated through correlation of adducts and isotopes in the aligned peak database, which also allows for their easy visualization and elimination as candidate biomarkers.

As described, twenty separate groups (4 doses × 5 time points) totaling 208 samples were analyzed. Following alignment and registration of the D-serine exposure data, more than 4000 isotopic peaks were originally registered and matched prior to automated deisotoping via MeDDL. During this process the isotopes were identified after peak matching was complete. The location ($\boldsymbol{M}_{avg}$,$\boldsymbol{T}_{avg}$) of each synthetic peak was used to initiate a search for monoisotopic peaks. For a given peak, a search was conducted to located monoisotopic peaks by looking for a peaks at location ($\boldsymbol{M}_{avg}+\mathbf{1},\boldsymbol{T}_{avg}$) ($\boldsymbol{M}_{avg}+\mathbf{2},\boldsymbol{T}_{avg}$) and ($\boldsymbol{M}_{avg}+\mathbf{3},\boldsymbol{T}_{avg}$). A match

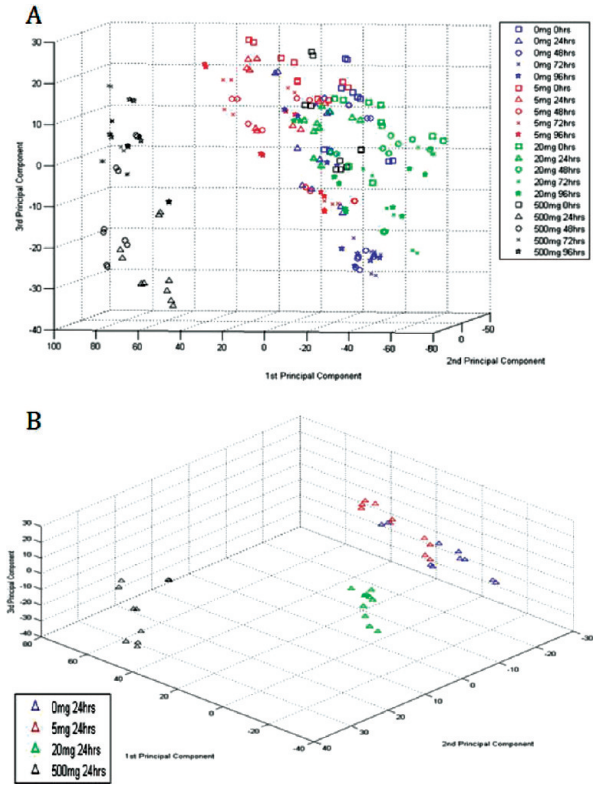


Figure 5. (A) Principal component analysis of LC−MS data for all experimental animal groups of the study. Legend is on the right of the figure. (B) Principal component analysis of LC−MS data for 0, 5, 20, and 500 mg/kg doses at 24 h only.

was found if a peak was located within the region defined by ($\boldsymbol{M}_{avg}+\mathbf{1}\pm\boldsymbol{M}_{\epsilon},\boldsymbol{T}_{avg}\pm\boldsymbol{T}_{\epsilon}$) where $\boldsymbol{M}_{\epsilon}$ and $\boldsymbol{T}_{\epsilon}$ is a user specified limit on mass and retention time variation between isotopic peaks. Once a potential isotope is identified, the intensity of the actual extracted peaks (main peak and isotopic peak) in each image is compared to verify that the isotopic peak has a decreasing level of intensity. If all peaks in the set and their corresponding isotopic peaks satisfy this requirement, then the isotopic peaks are tagged and can be hidden/removed by the user. A similar process is used to locate doubly and triply charge isotopic peaks and tag them for removal.

One of the novel aspects of the MeDDL peak alignment process is the use of a two stage process that begins with a rough peak match where only a few isolated peaks are identified between a reference image and each unregistered image. These initial peaks are used to compute a polynomial transformation between the reference image and the unregistered image producing a rough alignment. This is essentially a global process that handles systematic misalignment between images. In the peak matching phase, alignments are refined through a process similar to relaxation labeling.[16,17] After the rough alignment, a synthetic image is created by taking each image in turn and using every peak in the image as the center of a peak acceptance region. Any peak in any other image captured within the acceptance region is matched to this peak. The average mass and retention time ($\boldsymbol{M}_{avg}$, $\boldsymbol{T}_{avg}$) across the set of peaks is computed as ($\boldsymbol{M}_{avg}$, $\boldsymbol{T}_{avg}$).

(16) Rosenfeld, A.; Hummel, R.; Zucker, S. *IEEE Trans. Systems, Man Cybernet.* **1976**, *6*, 420–433.
(17) Wu, Q. *IEEE Trans Pattern Anal. Mach. Intel.* **1995**, *17*, 843–853.

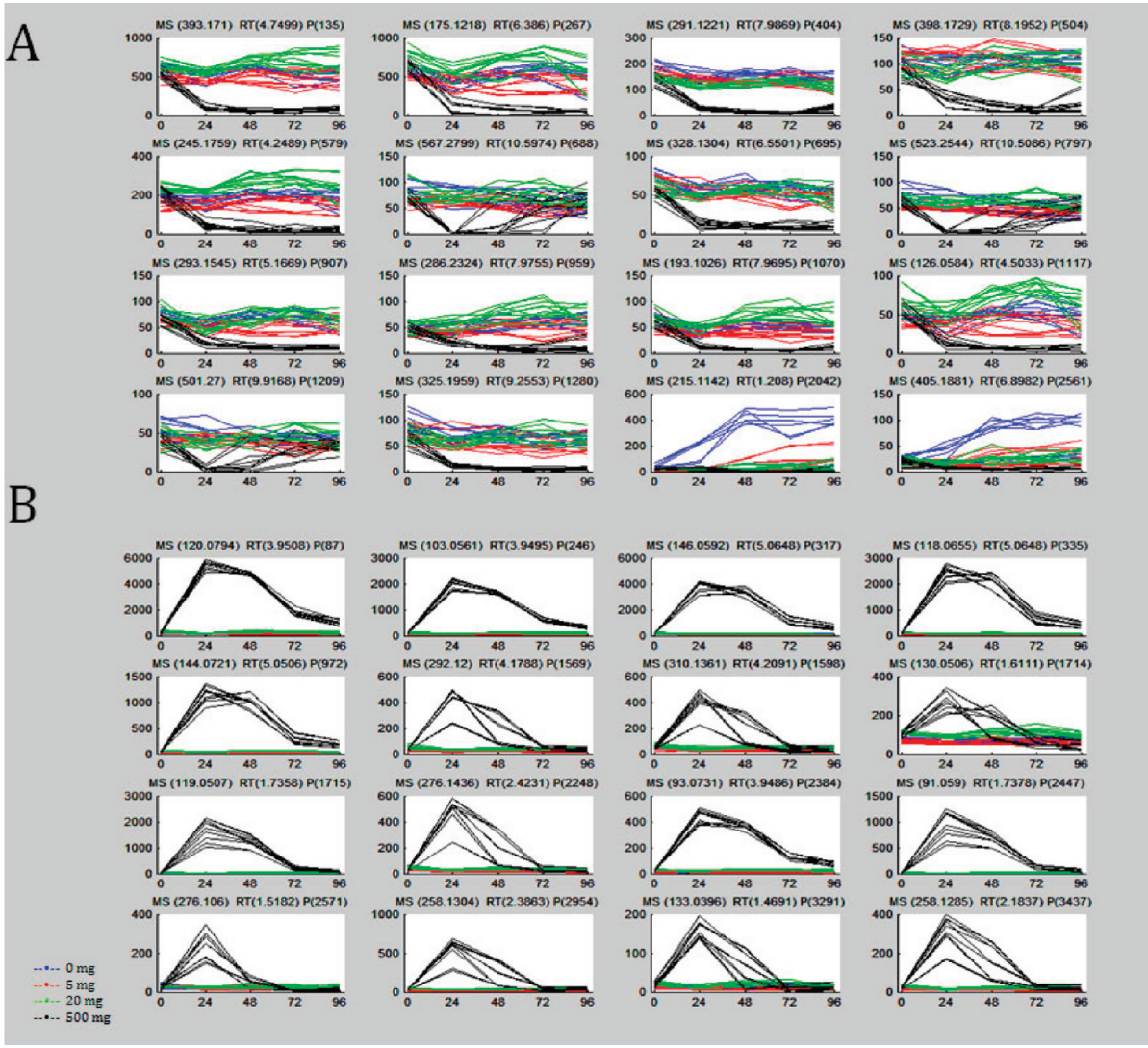


Figure 6. Examples of selected peak plots of negative (A) and positive (B) changes after 500 mg/kg D-Serine exposure.

Initially, the acceptance region is very small so peaks that were not well-aligned by the polynomial function will not be matched. The relaxation process slowly opens the size of the acceptance region to attempt to draw in one peak from each image. Each time a new peak is captured within the acceptance region, the average (M_{avg}, T_{avg}) is recalculated. Thus, the acceptance region gradually shifts and coalesces to maximize the number of matched peaks across all images at the final value of (M_{avg}, T_{avg}). This combination of global alignment/local refinement allows the matching to respond to both systematic misalignments as well outliers that appear as random variations within individual images. Additionally, although MeDDL accurately aligned all data generated by our laboratory, the retention time variation observed with the Waters Corp. Acquity UPLC was minimal (<0.25 min). As such, alignment was achieved through the use of a second order polynomial and our two stage peak alignment process. During development, our group evaluated use of higher order functions; however, we deemed it unnecessary for our use. This can be easily modified to be a user editable feature through the software interface if necessary for other chromatography systems.

During the analysis of the exposure data, two of the primary tools included in the MeDDL platform were utilized by our group, principal component analysis[18] (PCA), and a novel fold change filter. The design of the fold change filter analysis is based on a multilevel statistical model that views the behavioral response (intensity) of each synthetic peak as a normally distributed random with the added assumption that the behavior of peaks within individual images is correlated. On the basis of this underlying statistical model, the system is designed to handle longitudinal data sets consisting of subjects exposed to multiple levels of treatments. The statistical models are designed to allow the user to perform statistical tests for significant differences between treatment levels, significant differences between treatment time points, or significant differences between any combination of treatment levels or time points. In future applications, if other analysis tools become required, MeDDL is easily expandable through its use of the MVC software architecture, previously described. This software architecture allows a programmer to extend the functionality of the system as follows (1) add a new

(18) Richmond, B.; Optican, L.; Podell, M.; Spitzer, H. *J. Neurophysiol.* **1987**, *57*, 132.

choice to any pull-down menu in system menu, (2) install a new callback for the menu item that invokes a user defined function, and (3) create a new user function (userFunction.m). The user code added to userFunction.m has full access to all Matlab libraries (e.g., image processing, signal processing, pattern recognition, statistics, etc.) and full access to the summary data describing the matched peaks, the full description of every peak in a matched set and the raw data for every image. This allows a programmer/user to add new functionality to the system without altering the existing functionality.

PCA was performed for all groups of study animals and is shown in Figure 5. The PCA plot demonstrates clear separation between sample dosage and time groups with the majority of metabolomic changes in urine observed at 24, 48, and 72 h post treatment with 500 mg/kg D-Serine. The number of peaks that undergo at least a 2-fold change is 19× higher for the 500 mg/kg dose than the 5 mg/kg dose, with the changes literally disappearing at 96 h, most likely indicating kidney recovery.

At 24 h postdosing for the 500 mg/kg group, as many as 426 peaks show a greater than 2-fold change with the peak intensity cutoff set to a minimum of 100. Although the 2-fold increase was established based on our goals of identifying differential yet detectable metabolite biomarker profiles, the fold change filter incorporates detailed statistical analysis including ANOVA (1-way, 2-way, and *N*-way) among the selected subject groups. The user can optionally perform multiple pairwise comparison tests among the means of groups to determine whether or not all differences among group means satisfy a user defined level of significance. A Bonferroni correction is applied to compensate for the tendency to incorrectly find a single pairwise significant difference among multiple comparisons. Further, five metabolite peaks exceeded 100-fold change with the same intensity threshold. It is worthy to note that a number of peaks exhibit a statistically significant change while their intensities are relatively low, with most of these peaks demonstrating negative changes in our analysis. Examples of negative and positive changes are shown in Figure 6. We have excluded isotopic peaks in our data analysis; however, some percentage of differentiated peaks can be attributed to adduct acquisition by metabolites as well as water loss. Thus, the difference of 18 mass units between peaks 1569 and 1598; 952 and 246; 1642 and 1532; 1697 and 1664; and 3277 and 42 strongly suggest a water loss with each set of ions eluting from the column concurrently.

Metabolite identification is currently in progress with a list of potential metabolites shown in Table 2. Purchased metabolite standards were run under the same LC conditions followed by MS/MS as selected samples. Matching retention times and MS/MS fragmentation data generally indicate the conclusive identification of a metabolite, which has been demonstrated in a candidate biomarker identified in this study, 3-indolylacetic acid. The MS/MS spectrum of 3-indolylacetic acid along with a spectrum from rat urine samples is shown in Figure 7. This figure also demonstrates the ability of the MeDDL to automate spectral normalization. In MeDDL, normalization begins by computing the mean (m_1, m_2, m_3,...) and standard deviation (s_1, s_2, s_3,...) of the values of all subjects in each treatment group at the first time point. The peak intensity value $p_s(t)$ for each subject s in group j at time t is then normalized as follows: $(p_i(t) - m_j)/s_j$. This effectively shifts all the plots so the mean value of the first time point in each treatment group is zero (see Figure 7C).

Table 2. List of Potential Metabolites (Shown as *m/z*) Identified in Urine of Rats after 500 mg/kg D-Serine Exposure at 24 and 48 h after the Exposure[a]

m/z	retention time	treatment *x* time *p*-value
521.2412	9.071536	3.31×10^{-5}
523.2543	10.508634	8.04×10^{-5}
491.2421	7.5071826	1.44×10^{-4}
501.2701	9.916751	1.74×10^{-4}
714.1853	1.6265737	2.22×10^{-4}
611.3051	10.67593	2.60×10^{-4}
609.2873	9.332852	5.72×10^{-4}
567.2799	10.59735	7.30×10^{-4}
613.3242	11.733243	9.17×10^{-4}
383.1925	4.111277	0.00169446
779.352	1.4149536	0.00176583
290.1256	1.469668	0.00190476
655.3292	10.74838	0.00203854
435.159	10.100951	0.00356158
589.3217	10.138795	0.00407047
479.2284	10.405753	0.0055822
701.3726	11.777291	0.00567922
326.1946	9.491132	0.00787954
553.2582	4.0478706	0.01769215
633.3461	10.232473	0.01919617
212.1025	1.1299926	0.02197477
330.0621	2.3208911	0.02478929
533.1005	4.628623	0.04081653
511.2622	4.253147	0.04097449
290.1258	1.7111521	0.04116634

[a] Fold change filter set at 10 fold and higher.

CONCLUSIONS

The data clearly demonstrate dramatic changes in the urinary metabolic profile in response to the kidney toxicant, D-Serine. A list of potential metabolites corresponding to masses identified in urine of rats is presented. D-Serine metabolomic profiling demonstrates that most changes occur between 24−72 h. The most dramatic changes occur at the 24-h time point after exposure to 500 mg/kg D-Serine. The data suggests that near-normal kidney function resumes at 96 h. Metabolite identification of selected peaks from the study is currently in progress.

Although the ability to visualize the experiment at all levels may constitute the authors' ideal for biomarker discovery and differential metabolite analysis, we feel it adds considerably to this effort by allowing the user to differentiate metabolite profiles in a large time-dose study while maintaining the ability to focus on individual metabolites and spectra for subsequent identification. The overall framework was rapidly prototyped using MathWork's Matlab software language and is being translated to the general purpose, platform independent language, Python, to support wide dissemination of the tool. The MeDDL tool dramatically reduced manpower costs in our research by providing scaffolding for the rapid development and verification of new algorithms without the need to create a large amount of supporting software. MeDDL also offered the potential for staff scientists to visualize data in new ways, providing novel insight into the experimental results and facilitating metabolomic biomarker discovery. It should be

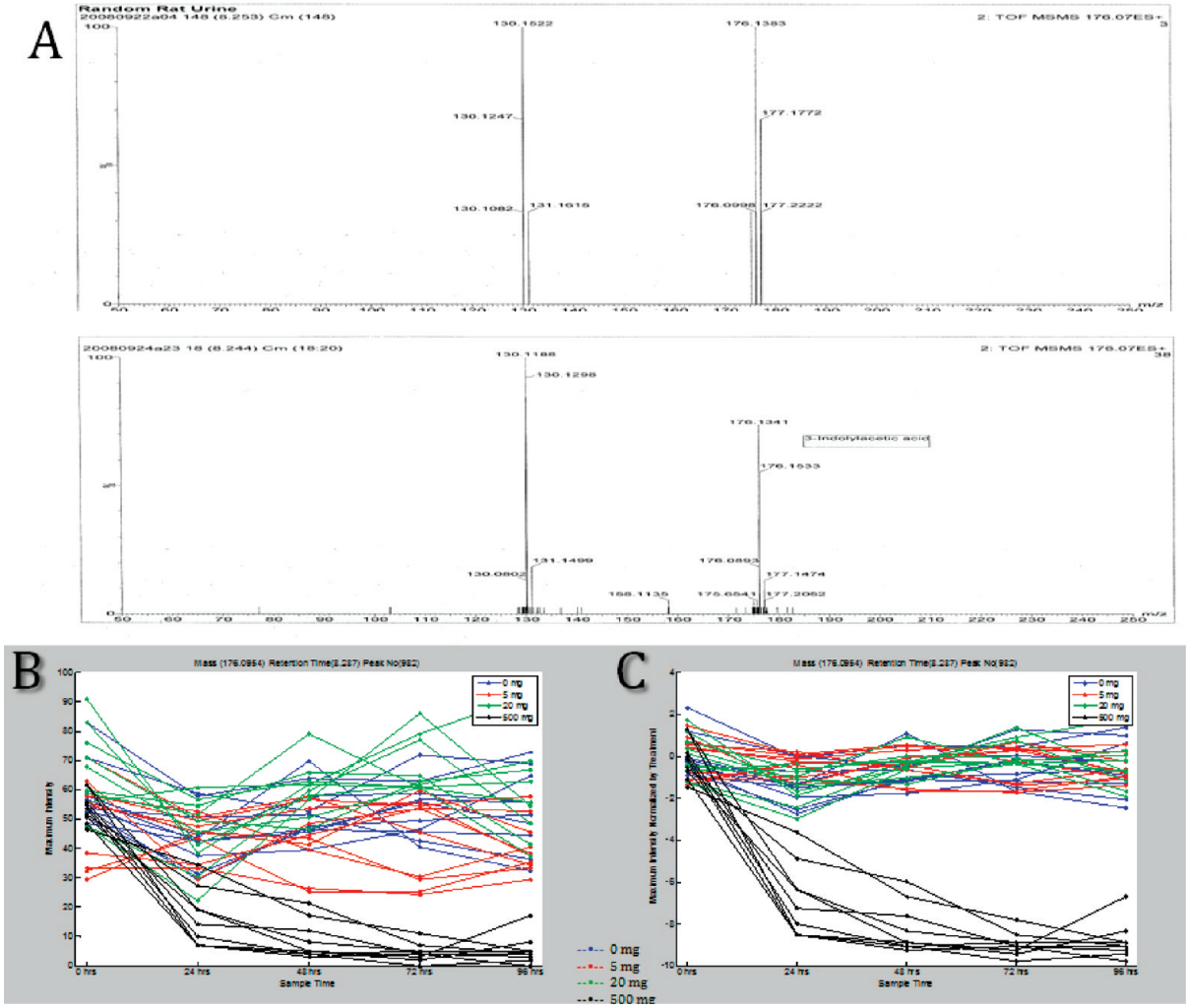


Figure 7. Identification of a selected metabolite. Retention times and MS/MS fragmentation of 3-Indolylacetic acid is shown for standards along with corresponding ms/ms from D-Serine urine samples in A. Not normalized (B) and normalized (C) plots show changes for 3-Indolylacetic acid throughout the course of D-Serine study.

noted that a number of tools have recently been proposed in the literature, which show great advancements in metabolomic and LC−MS analysis capability.[19−24] However, the MeDDL tool, through its emphasis on visualization, provides unique opportunities by combining the following: easy use of both GC−MS and LC−MS data; use of both manufacturer specific data files as well as netCDF (network Common Data Form); preprocessing (peak registration and alignment in both time and mass); powerful visualization tools; and built in data analysis functionality.

(19) Katajamaa, M.; Oresic, M. *BMC Bioinformatics* **2005**, *6*, 179.
(20) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.
(21) Baran, R.; Kochi, H.; Saito, N.; Suematsu, M.; Soga, T.; Nishioka, T.; Robert, M.; Tomita, M. *BMC Bioinformatics* **2006**, *7*, 530.
(22) Bunk, B.; Kucklick, M.; Jonas, R.; Munch, R.; Schobert, M.; Jahn, D.; Hiller, K. *Bioinformatics* **2006**, *22*, 2962–2965.
(23) Broeckling, C. D.; Reddy, I. R.; Duran, A. L.; Zhao, X.; Sumner, L. W. *Anal. Chem.* **2006**, *78*, 4334–4341.
(24) Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J. *Bioinformatics* **2008**, *24*, 732–737.

ACKNOWLEDGMENT

This bioinformatics analysis platform was developed in support of Department of Defense Joint Services Defense Technology Objective MD.34, toward the identification of common stress biomarkers and low-level effects of toxic exposure. We would like to extend special thanks to Deirdre Mahle, Nicholas DelRaso, Tyler Bennett, Mitchell Meade, and Lining Qi for their generous help and support in completion of this work. Finally, this project would not have been possible without the vision, enthusiasm, and support of Dr. John J. Schlager toward the application of LC/MS in biomarker discovery. His significant contributions to this effort are greatly appreciated.

Received for review January 6, 2010. Accepted April 20, 2010.

AC100034U